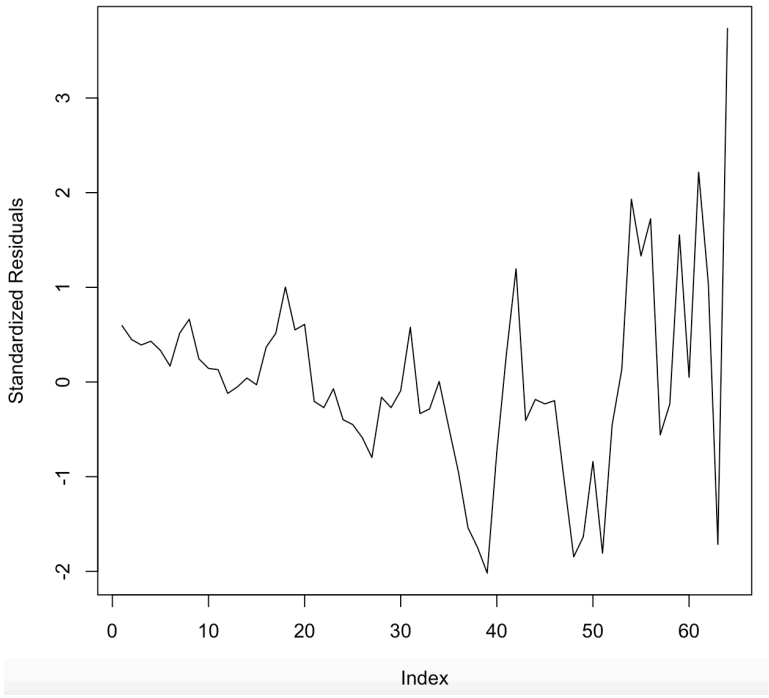
Homework 2.1

Exercise 3.7 The data file winnebago contains monthly unit sales of recreational vehicles from Winnebago, Inc., from November 1966 through February 1972.

(a) Display and interpret the time series plot for these data.

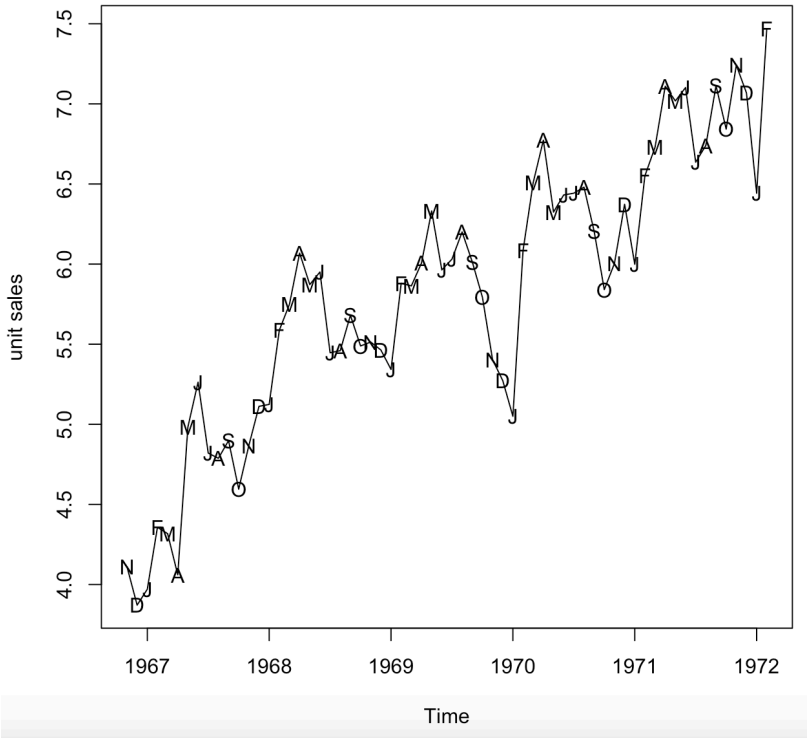
The moving average and the variance of unit sales appear to both increase with time. There appears to be seasonality in the raw data with dips in unit sales occurring in similar months (October-January) throughout the years.

(b) Use least squares to fit a line to these data. Interpret the regression output. Plot the standardized residuals from the fit as a time series. Interpret the plot.



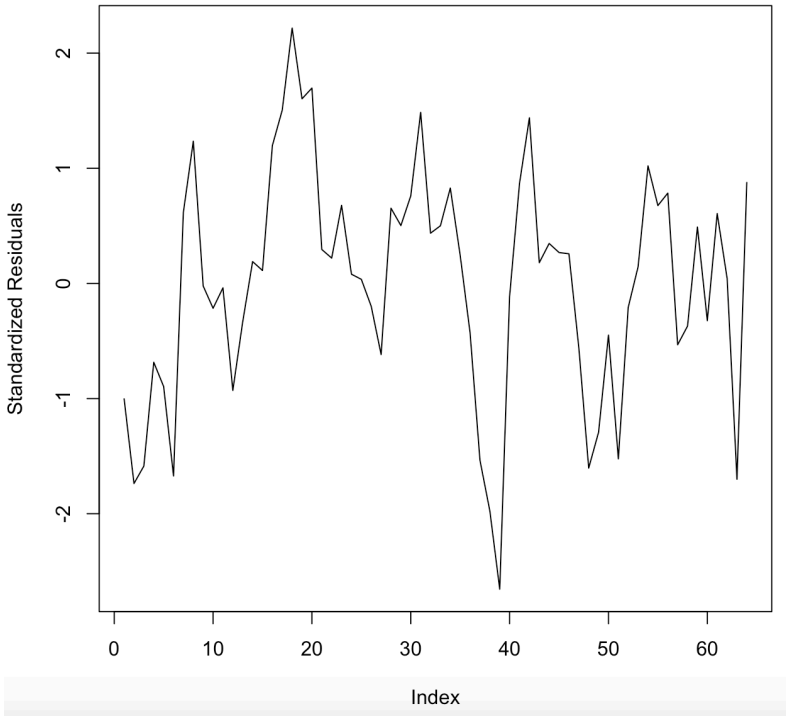
The homoscedasticity assumption that is necessary for linear regression models is clearly violated here. The variance of the data fans out as the time is increased.

(c) Now take natural logarithms of the monthly sales figures and display and interpret the time series plot of the transformed values.



The log function brings down the skew that is in the later end of the time series data. The moving average now looks more linear and less like an exponential increase as it appeared in the raw data.

(d) Use least squares to fit a line to the logged data. Display and interpret the time series plot of the standardized residuals from this fit.



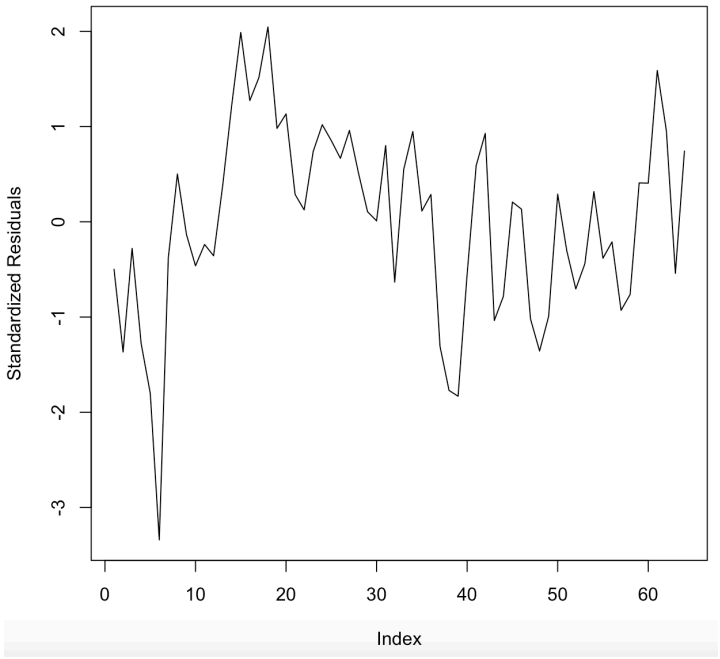
The homoscedasticity assumption looks much improved here in this residual plot. The residuals do not appear to have any pattern and there are not any residuals with an absolute value greater than 3, which is a good indicator of normality as well.

(e) Now use least squares to fit a seasonal-means plus linear time trend to the logged sales time series and save the standardized residuals for further analysis. Check the statistical significance of each of the regression coefficients in the model.

```
Call:
lm(formula = log.winnebago ~ month + time(log.winnebago))

Residuals:
     Min       1Q   Median       3Q      Max
-0.92501 -0.16328  0.03344  0.20757  0.57388

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -997.33061   50.63995 -19.695  < 2e-16 ***
monthFebruary          0.62445    0.18182   3.434 0.001188 **
monthMarch             0.68220    0.19088   3.574 0.000779 ***
monthApril             0.80959    0.19079   4.243 9.30e-05 ***
monthMay               0.86953    0.19073   4.559 3.25e-05 ***
monthJune              0.86309    0.19070   4.526 3.63e-05 ***
monthJuly              0.55392    0.19069   2.905 0.005420 **
monthAugust            0.56989    0.19070   2.988 0.004305 **
monthSeptember         0.57572    0.19073   3.018 0.003960 **
monthOctober           0.26349    0.19079   1.381 0.173300
monthNovember          0.28682    0.18186   1.577 0.120946
monthDecember          0.24802    0.18182   1.364 0.178532
time(log.winnebago)    0.50909    0.02571  19.800  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3149 on 51 degrees of freedom
Multiple R-squared:  0.8946,    Adjusted R-squared:  0.8699
F-statistic: 36.09 on 12 and 51 DF,  p-value: < 2.2e-16
```

All of the months other than October, November, and December are showing up as having a statistically significant difference compared to the month of January. The late spring and summer months appear to have the greatest difference when compared to January.

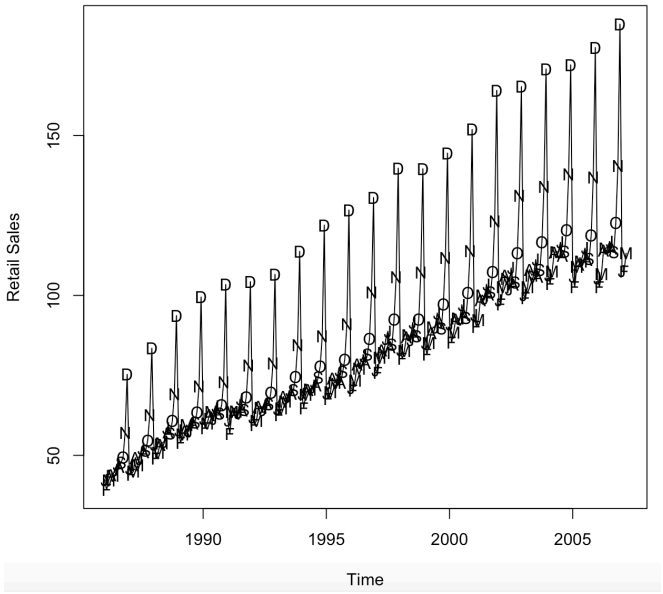
(f) Display the time series plot of the standardized residuals obtained in part (e). Interpret the plot.



The standardized residuals don't appear to show any patterns. The variance stays consistent throughout so the homoscedasticity assumption is met. There is a spike that goes beyond the absolute value of 3 rule of thumb for normalized data, however I don't believe that this temporary spike is enough to rule out this model.

Exercise 3.8 The data file retail lists total U.K. (United Kingdom) retail sales (in billions of pounds) from January 1986 through March 2007. The data are not “seasonally adjusted,” and year 2000 = 100 is the base year.

(a) Display and interpret the time series plot for these data. Be sure to use plotting symbols that permit you to look for seasonality.



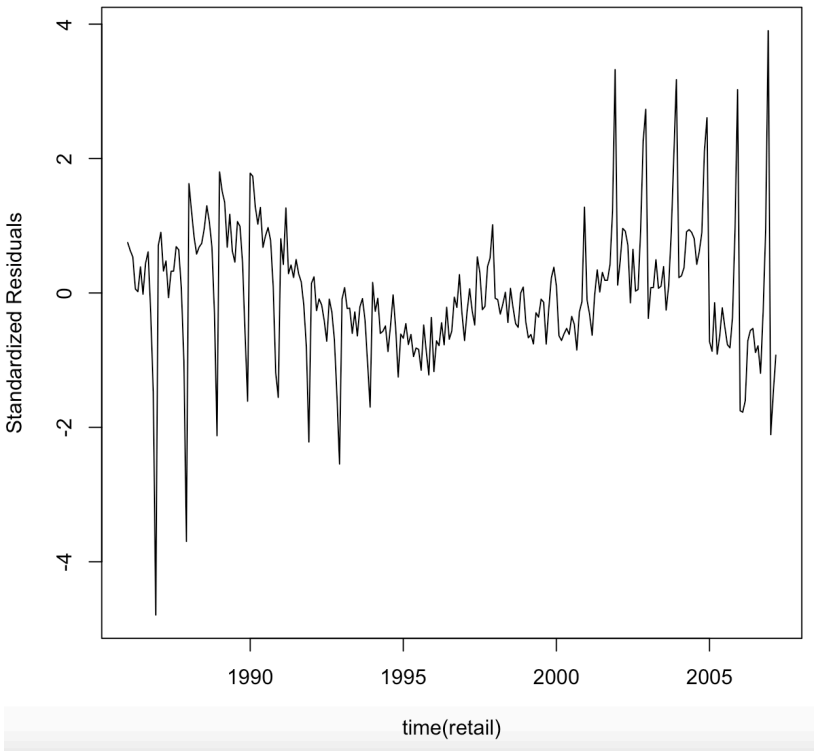
The data is a little bit convoluted to see most months. It is clear to see from this plot however that December and November stand out as far as retail sales go.

(b) Use least squares to fit a seasonal-means plus linear time trend to this time series. Interpret the regression output and save the standardized residuals from the fit for further analysis.

```
Call:
lm(formula = retail ~ month + time(log.retail))

Residuals:
     Min       1Q   Median       3Q      Max
-19.8950  -2.4440  -0.3518   2.1971  16.2045

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -7.249e+03  8.724e+01 -83.099  < 2e-16 ***
monthFebruary    -3.015e+00  1.290e+00  -2.337  0.02024 *
monthMarch        7.469e-02  1.290e+00   0.058  0.95387
monthApril        3.447e+00  1.305e+00   2.641  0.00880 **
monthMay          3.108e+00  1.305e+00   2.381  0.01803 *
monthJune         3.074e+00  1.305e+00   2.355  0.01932 *
monthJuly         6.053e+00  1.305e+00   4.638 5.76e-06 ***
monthAugust       3.138e+00  1.305e+00   2.404  0.01695 *
monthSeptember    3.428e+00  1.305e+00   2.626  0.00919 **
monthOctober      8.555e+00  1.305e+00   6.555 3.34e-10 ***
monthNovember     2.082e+01  1.305e+00  15.948  < 2e-16 ***
monthDecember     5.254e+01  1.305e+00  40.255  < 2e-16 ***
time(log.retail)  3.670e+00  4.369e-02  83.995  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.278 on 242 degrees of freedom
Multiple R-squared:  0.9767,    Adjusted R-squared:  0.9755
F-statistic:   845 on 12 and 242 DF,  p-value: < 2.2e-16
```

It is shown that all months other than March contain retail sales that are significantly different from January in this model. The model appears to fit the data very well as the Adjusted R-squared value is 0.9755, indicating that the explanatory variables season and time explain away approximately 97.55% of the variance in the data.

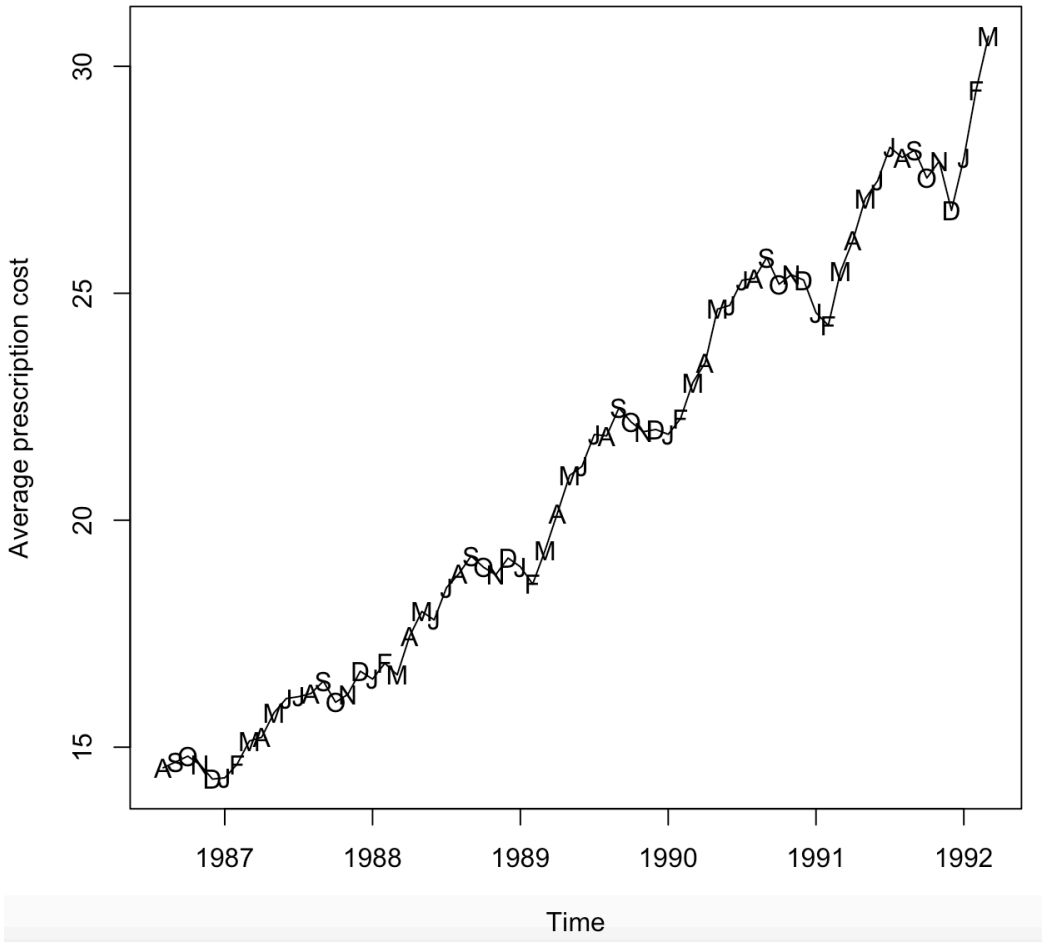
(c) Construct and interpret the time series plot of the standardized residuals from part (b). Be sure to use proper plotting symbols to check on seasonality.



The standardized residuals do appear to have some patterning. There are more outliers in the negative direction in early times whereas there are more outliers in the positive direction on the later end of the data. The moving average trend also appears like it could have slightly quadratic curvature to it.

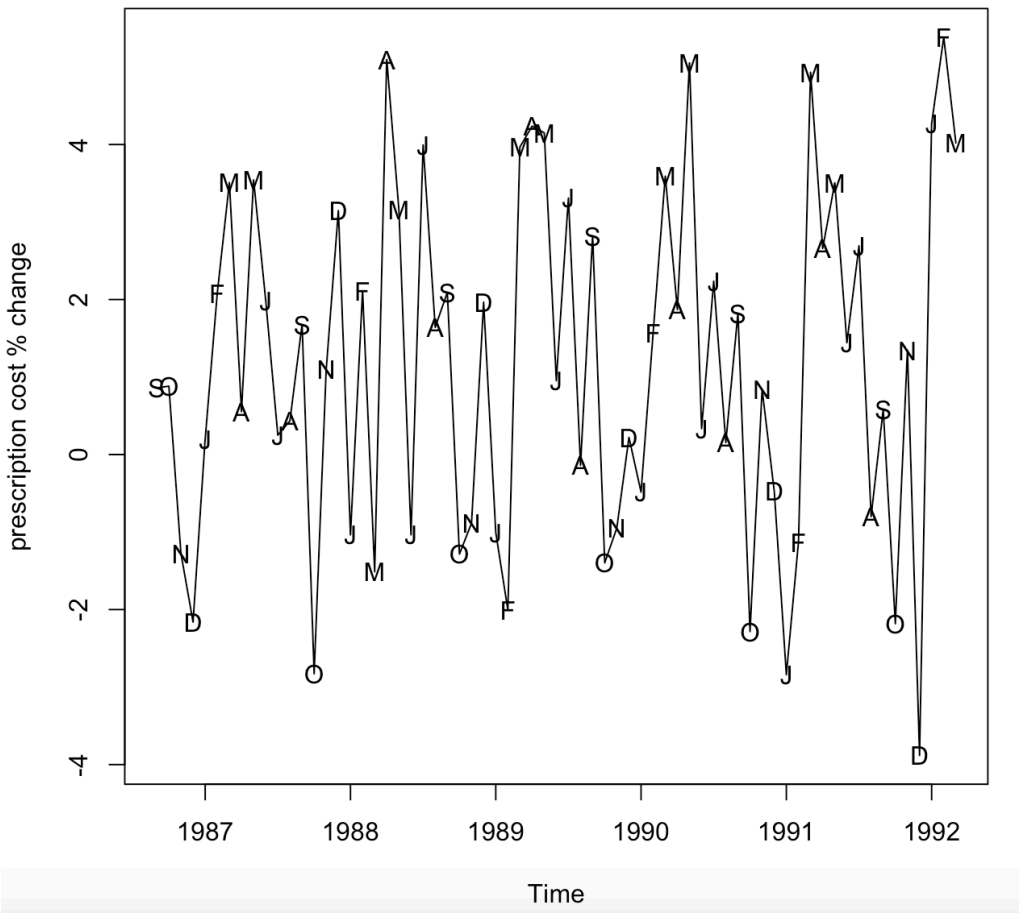
Exercise 3.9 The data file prescrip gives monthly U.S. prescription costs for the months August 1986 to March 1992. These data are from the State of New Jersey's Prescription Drug Program and are the cost per prescription claim.

(a) Display and interpret the time series plot for these data. Use plotting symbols that permit you to look for seasonality.



You can see a dip in Monthly U.S. average prescription costs in the winter months (December, January, February) whereas the prices climb back in the summer months. The average costs appear to be continuously increasing throughout all of the years in this dataset in a linear fashion.

(b) Calculate and plot the sequence of month-to-month percentage changes in the prescription costs. Again, use plotting symbols that permit you to look for seasonality.



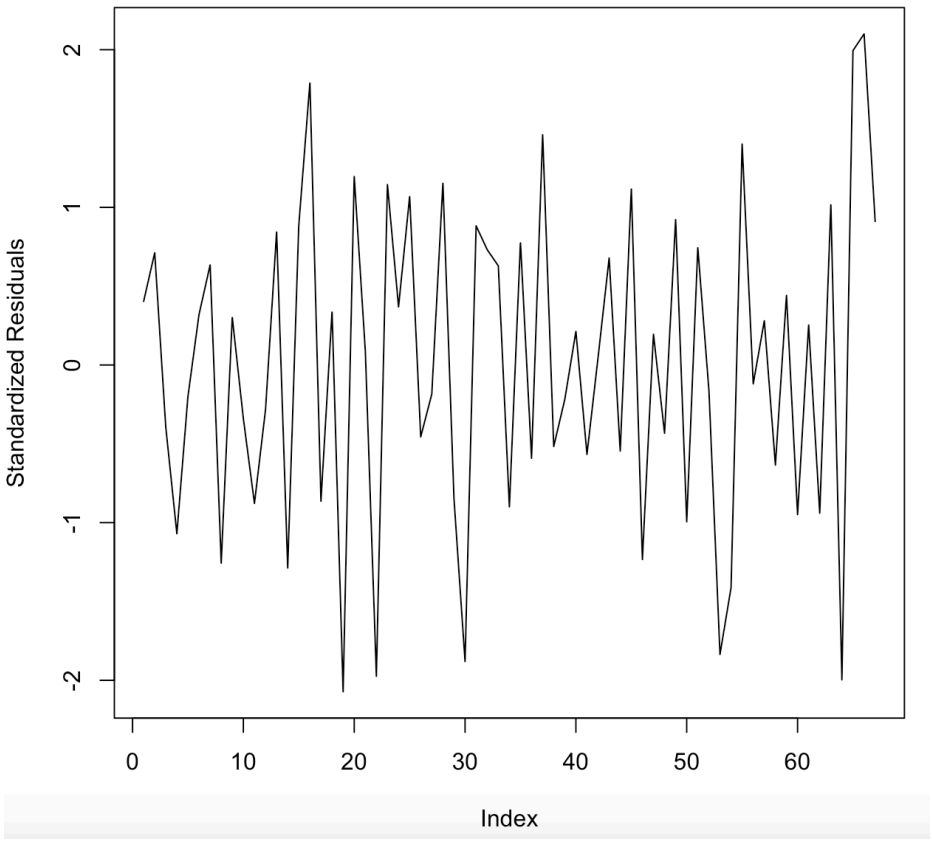
There appears to be more positive percent month to month difference days than negative month to month percent difference days. The data appears fairly sinusoidal in nature. The pattern appears to be that later months of the year decrease in percent change of prescription costs while the middle months of the year have percent increase changes.

(c) Use least squares to fit a cosine trend with fundamental frequency 1/12 to the percentage change series. Interpret the regression output. Save the standardized residuals.

```
Call:
lm(formula = prescrip_percent_change ~ har)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8444 -1.3742  0.1697  1.4069  3.8980

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.2217     0.2325   5.254 1.82e-06 ***
harcos(2*pi*t)  -0.6538     0.3298  -1.982   0.0518 .
harsin(2*pi*t)   1.6596     0.3269   5.077 3.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.897 on 64 degrees of freedom
Multiple R-squared:  0.3148,    Adjusted R-squared:  0.2933
F-statistic:  14.7 on 2 and 64 DF,  p-value: 5.584e-06
```

The Cosine trend does not appear to be significantly different from 0 as its p value is greater than 0.05. The Sine trend function does however appear to have a coefficient significantly different than 0. The r squared value is not very large, indicating that this model does not do the best job at describing the variation in the data.

(d) Plot the sequence of standardized residuals to investigate the adequacy of the cosine trend model. Interpret the plot.



The standardized residuals appear to be have consistent variance throughout the data. The standardized residuals points all fall within approximately 2 standard deviations indicating normality. These point to the fact that this model is a decent fit for the data.

Exercise 3.13 (Continuation of Exercise 3.7) Return to the winnebago time series.

(a) Calculate the least squares residuals from a seasonal-means plus linear time trend model on the logarithms of the sales time series.

1	2	3	4	5	6
-0.498578789	-1.366906868	-0.279469640	-1.280415130	-1.798865046	-3.340284769
7	8	9	10	11	12
-0.376314763	0.501524449	-0.132805716	-0.462068655	-0.237848190	-0.356701604
13	14	15	16	17	18
0.378678738	1.224079643	1.987847009	1.273179866	1.514026564	2.046024616
19	20	21	22	23	24
0.981193796	1.132853837	0.287319541	0.124944466	0.739350894	1.019477922
25	26	27	28	29	30
0.852001514	0.666679673	0.959528214	0.511472806	0.105836780	0.009645785
31	32	33	34	35	36
0.800623990	-0.633401748	0.551894479	0.947519293	0.113007069	0.287346926
37	38	39	40	41	42
-1.306359561	-1.770532710	-1.831627818	-0.551002893	0.591327746	0.928230166
43	44	45	46	47	48
-1.036519631	-0.783437501	0.207542067	0.133072829	-1.021150823	-1.356676439
49	50	51	52	53	54
-0.993583007	0.290542254	-0.301956359	-0.704694867	-0.441185980	0.318004614
55	56	57	58	59	60
-0.381959409	-0.211200346	-0.929743374	-0.762981151	0.408329869	0.405973419
61	62	63	64		
1.589448694	0.958383047	-0.540719684	0.742951380		

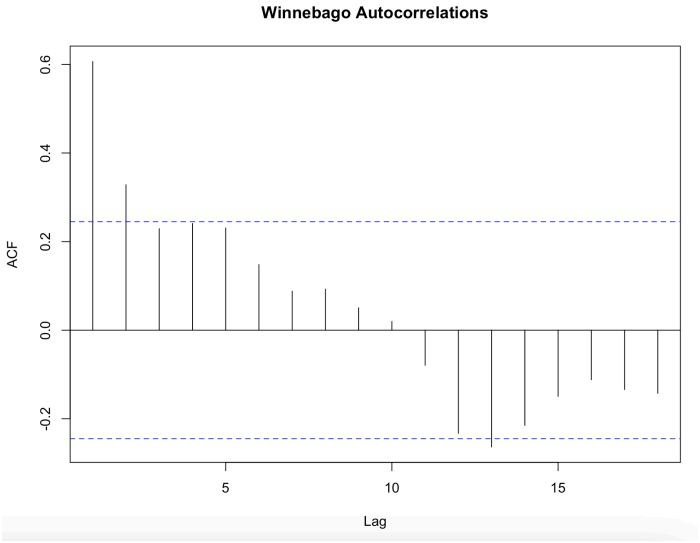
The standardized residuals of the seasonal-means plus linear time trend model on the logarithms of the Winnebago sales time series can be seen above.

(b) Perform a runs test on the standardized residuals and interpret the results.

```
$pvalue
[1] 0.000243

$observed.runs
[1] 18

$expected.runs
[1] 32.71875

$n1
[1] 29

$n2
[1] 35

$k
[1] 0
```

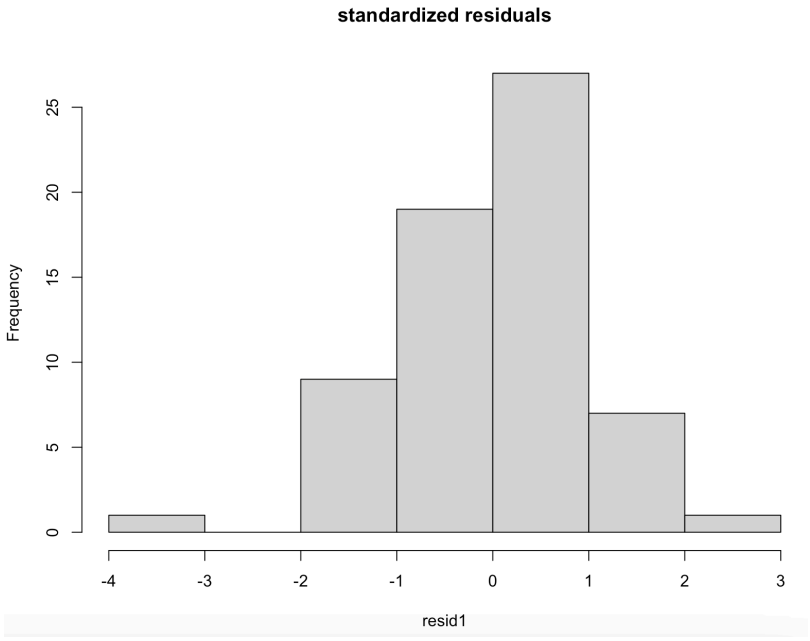
As can be seen from the image above the expected number of runs is greater than the number of runs observed. This indicates that there is likely a dependence on time for this data. The p value supports this lack of independence of time because it is lower than the statistical significance threshold of 0.05.

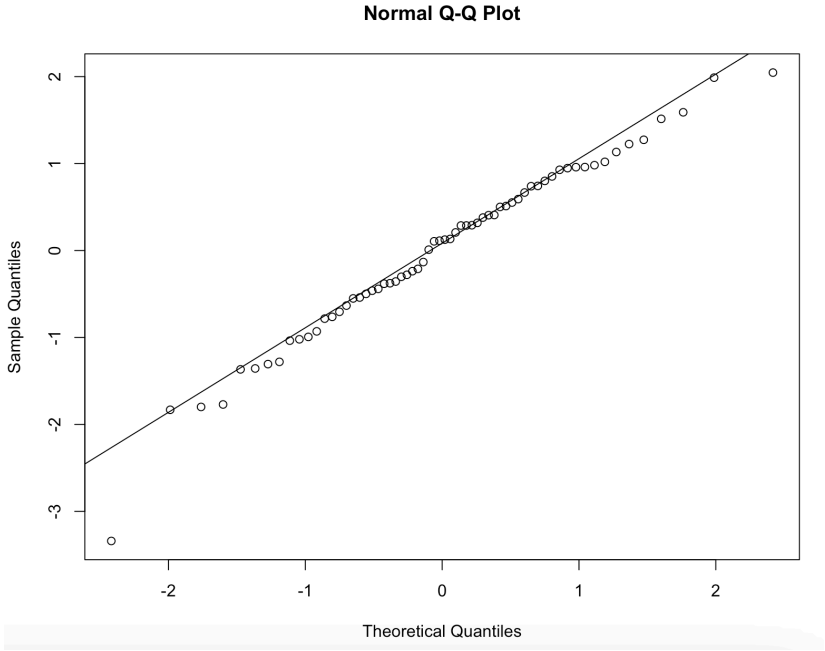
(c) Calculate and interpret the sample autocorrelations for the standardized residuals.



The sample autocorrelations plot shows multiple lag points with autocorrelation magnitudes above a threshold for concern. This is an indication that the data is not independent of time.

(d) Investigate the normality of the standardized residuals (error terms). Consider histograms and normal probability plots. Interpret the plots.





The histogram of the standardized residuals show a nice bell curve shape. The qq plot shows residuals mostly fall on the line that would anticipate normal residuals. There is one outlier however on the negative side of the data, a value was much lower than it was anticipated to be given a normal dataset. Other than this one problematic data point though this is a good sign of normality in the data for this model.

Exercise 3.14 (Continuation of Exercise 3.8) The data file retail contains U.K. monthly retail sales figures

(a) Obtain the least squares residuals from a seasonal-means plus linear time trend model.

1	2	3
0.7492119807	0.6310464807	0.5358576056
4	5	6
0.0596999580	0.0195587238	0.3877111867
7	8	9
-0.0182887257	0.4416151299	0.6113552062
10	11	12
-0.2878084416	-1.4599324824	-4.7916549274
13	14	15
0.7075033931	0.9020425758	0.3260754451
16	17	18
0.4758594503	-0.0694879054	0.3223372955
19	20	21
0.3257743587	0.6889573666	0.6419841700
22	23	24
0.0805971862	-0.9940578970	-3.6967352331
25	26	27
1.6267737977	1.2206120138	0.8373787177
28	29	30
0.5787655059	0.6829252940	0.7378667207
31	32	33
0.9576324276	1.2964378924	1.0572137636
34	35	36
0.6886483593	-0.3128880653	-2.1236659210

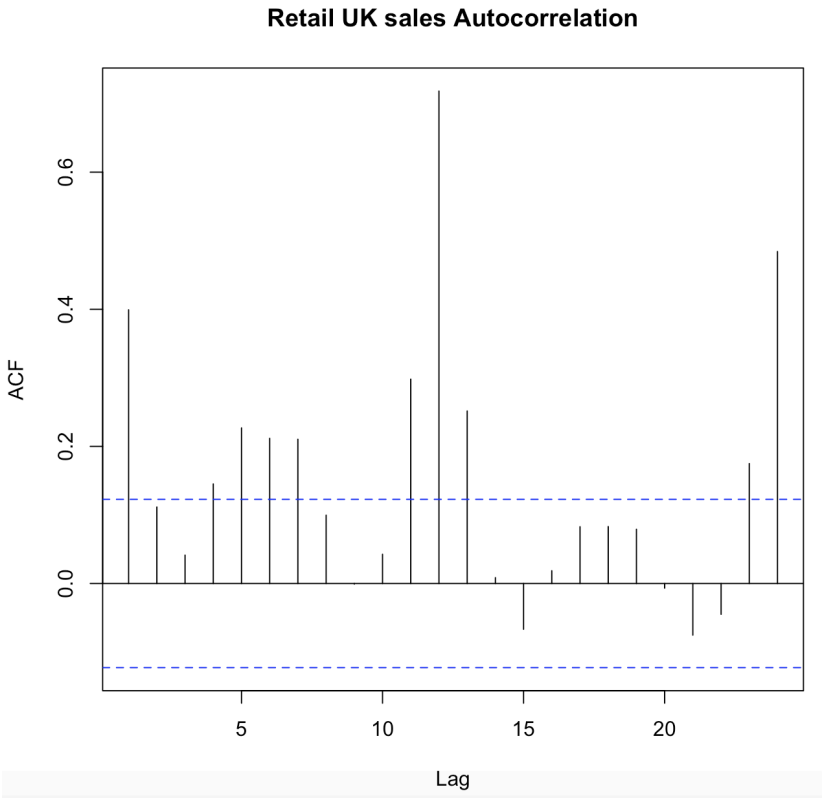
An image of some of the standardized residuals collected from the seasonal-means plus linear time trend model of the UK monthly retail sales.

(b) Perform a runs test on the standardized residuals and interpret the results.

```
$pvalue
[1] 9.19e-23

$observed.runs
[1] 52

$expected.runs
[1] 127.9333

$n1
[1] 136

$n2
[1] 119

$k
[1] 0
```

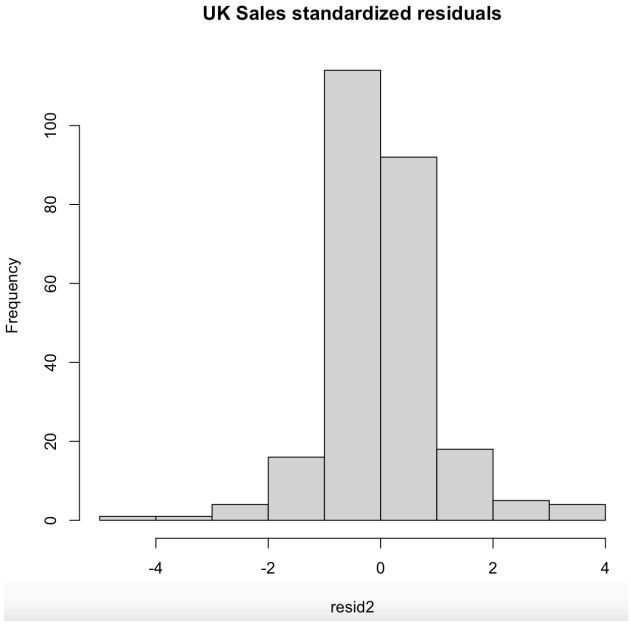
As can be seen from the image above the expected number of runs is far greater than the number of runs actually observed. This indicates that there is likely a dependence on time for this data. The p value further supports this lack of independence because it is much lower than the statistical significance threshold of 0.05.

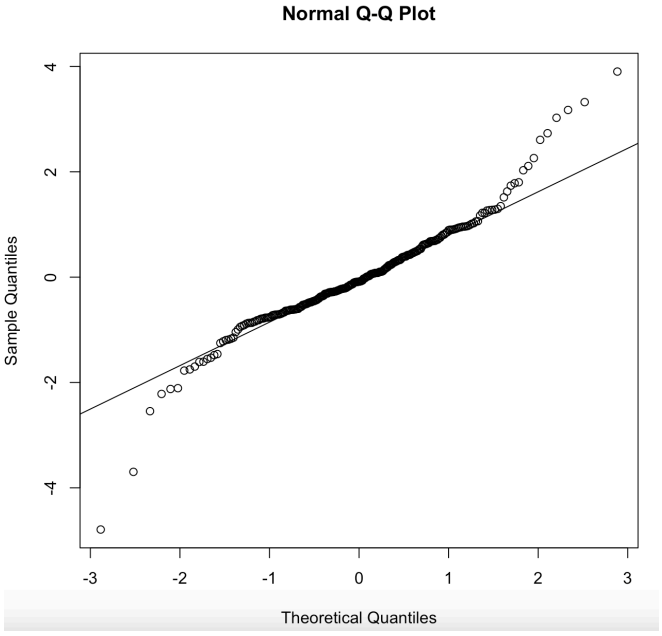
(c) Calculate and interpret the sample autocorrelations for the standardized residuals.



The sample autocorrelations plot shows multiple lag points with autocorrelation magnitudes above a threshold for concern. This is an indication that the data is not independent of time.

(d) Investigate the normality of the standardized residuals (error terms). Consider histograms and normal probability plots. Interpret the plots.





The histogram of the standardized residuals show a nice bell curve shape. The qq plot however highlights heavy tails. There are residuals on both extreme ends that fall outside the line that anticipate normal residuals. Normality would be a strong assumption for this model. A distribution that allows heavier tails such as t-distribution may be more appropriate.

Exercise 3.15 (Continuation of Exercise 3.9) Consider again the prescrip time series.

(a) Save the standardized residuals from a least squares fit of a cosine trend with fundamental frequency 1/12 to the percentage change time series.

```
         1          2          3          4          5          6          7          8
 0.40328878 0.71147920 -0.39591253 -1.06998646 -0.19927554 0.32090425 0.63445484 -1.25634312
         9         10         11         12         13         14         15         16
 0.30081792 -0.34064071 -0.87808147 -0.28261862 0.84324520 -1.28734899 0.88535150 1.78859210
        17         18         19         20         21         22         23         24
-0.86405986 0.33536365 -2.07283202 1.19530357 0.09166211 -1.97441741 1.14434009 0.36884409
        25         26         27         28         29         30         31         32
 1.06802235 -0.45590493 -0.18437205 1.15244230 -0.84764608 -1.88100709 0.88265927 0.73150076
        33         34         35         36         37         38         39         40
 0.62933710 -0.89943371 0.77426417 -0.59107892 1.46024238 -0.51681364 -0.21774166 0.21308837
        41         42         43         44         45         46         47         48
-0.56710841 0.04681480 0.67873583 -0.54551457 1.11621479 -1.23363818 0.19487851 -0.43300517
        49         50         51         52         53         54         55         56
 0.92205309 -0.99442782 0.74355421 -0.16046716 -1.83600586 -1.41203023 1.40155826 -0.11957089
```

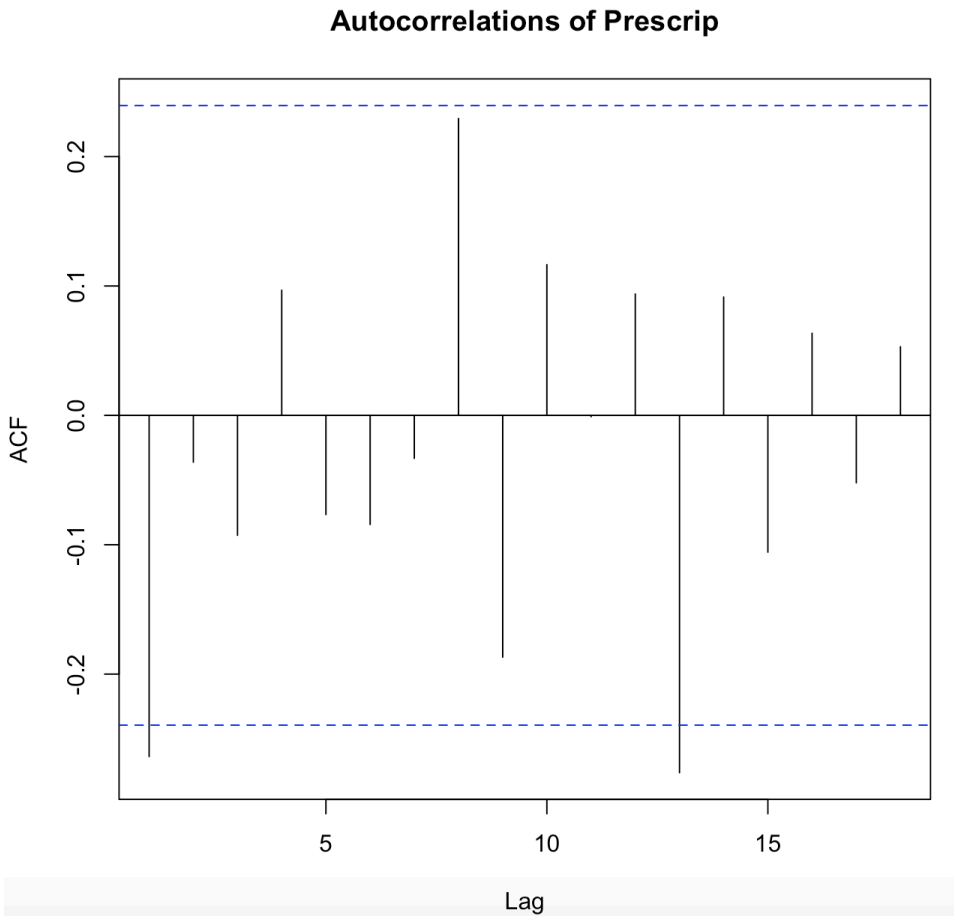
The standardized residuals of the seasonal-means plus linear time trend model on the logarithms of the prescrip sales time series can be seen above.

(b) Perform a runs test on the standardized residuals and interpret the results.

```
> runs(resid3)
$pvalue
[1] 0.0026

$observed.runs
[1] 47

$expected.runs
[1] 34.43284

$n1
[1] 32

$n2
[1] 35

$k
[1] 0
```

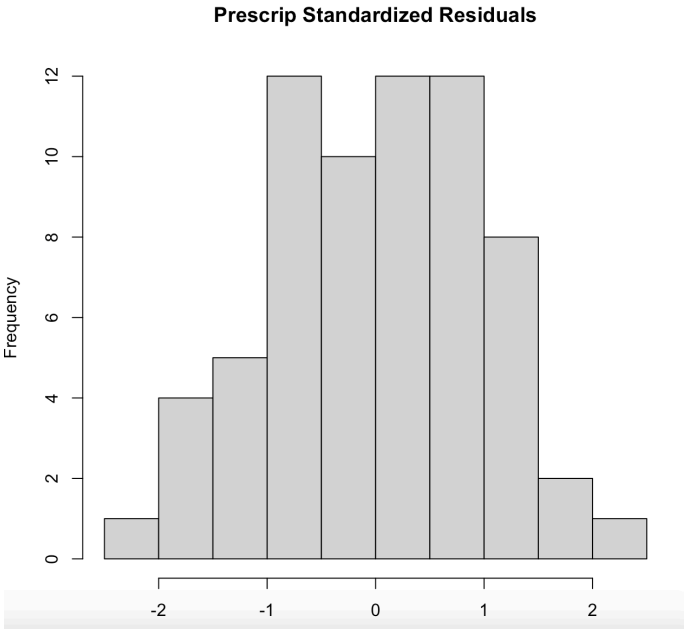
As can be seen from the image above the expected number of runs is far less than the number of runs actually observed. This indicates that there is likely a dependence on time for this data. The p value further supports this lack of independence because it is much lower than the statistical significance threshold of 0.05.

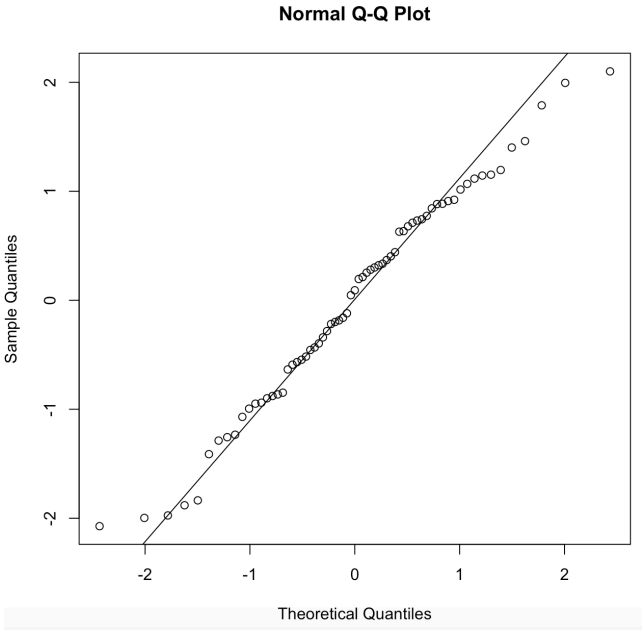
(c) Calculate and interpret the sample autocorrelations for the standardized residuals.



The sample autocorrelations plot shows multiple lag points with autocorrelation magnitudes above a threshold for concern. This is an indication that the data is not independent of time.

(d) Investigate the normality of the standardized residuals (error terms). Consider histograms and normal probability plots. Interpret the plots





The histogram of the standardized residuals show a nice bell curve shape. The qq plot highlights slightly lighter tails. The residuals are not as extreme as one would anticipate for normal residuals. This is not a big enough concern however and I believe normality is a good assumption for this model.